How to Derive Force Field Parameters by Genetic Algorithms: Modelling *tripod*-$Mo(CO)_3$ Compounds as an Example

Johannes Hunger, Stefan Beyreuther, Gottfried Huttner*, Kurt Allinger, Uwe Radelof, and Laszlo Zsolnai

Anorganisch-Chemisches Institut der Universität Heidelberg,
Im Neuenheimer Feld 270, D-69120 Heidelberg, Germany
Tel. (internat.): +49 (0)6221/549446
Fax (internat.): +49 (0)6221/545707
E-mail: GA@indi.aci.uni-heidelberg.de

Received October 28, 1997

Keywords: Genetic Algorithms / Force field calculations / Refinement of force field parameters / Tripod metal compounds / Conformational analysis

Force field parameters used to describe the conformation of coordination compounds involving transition metals are generally derived by a trial-and-error procedure, until a somehow satisfying agreement between the calculated and observed conformations of a few members of a class of related compounds is reached. It is shown in this paper that a more general and less biased alternative is available, applicable to many structures at a time. Genetic Algorithms will effectively optimize force field parameters in an automatic way, on the basis of a potentially exhaustive set of all the structural data available for a given class of compounds. The feasibility of this procedure has been demonstrated by the derivation of force field parameters describing the conformational behaviour of *tripod*-$Mo(CO)_3$ compounds [*tripod* = $RCH_2C(CH_2X)(CH_2Y)(CH_2Z)$, X,Y,Z = PR′R′′)] by simultaneous optimization based on the structure of ten individual molecules. With the force field parameters relevant to the organic part of these compounds taken from MM2*, the parameters involving contributions from the Mo center were refined. The agreement between observed and calculated structures is characterized by an rms deviation of around 0.3 Å for the ten structures contained in the data base. To assess the validity of this approach, the conformational space of $CH_3C(CH_2PPh_2)_3Mo(CO)_3$ was explored exhaustively. A contour diagram representing the relative energy of the molecule with respect to the rotational positions of its phenyl groups was found to effectively reproduce the scatter of these conformational parameters as earlier derived from an analysis of 82 relevant compounds. − As a further assessment, the conformational space of $CH_3C[CH_2P(o\text{-Tol})_2]_3Mo(CO)_3$, which was not included in the data base, has been analyzed. It is found that the structure corresponding to the global energy minimum corresponds to that observed in the crystal with an rms deviation of only 0.3 Å. The novel approach to problems of this type − Genetic Algorithms had not previously been applied in this context − thus appears promising.

Introduction

The reactivity of ligand-metal templates is largely determined by their conformations. Homogeneous catalysis by such templates has been shown to be strongly dependent on even minor changes in the conformation as well as in the conformational flexibility of the templates themselves.[1] Methods of predicting these conformational properties would hence be of considerable practical importance. "Understanding" conformations in the sense of being able to model them would be a crucial step in any attempt to understand homogeneous catalysis. In spite of the success quantum chemical methods nowadays have to rationalize and predict the properties of even relatively large ligand−metal templates[2][3], these methods are not yet in a state to allow for an extensive conformational search in a practicable amount of time. As an alternative, force field methods have become popular since N. L. Allinger and his group demonstrated how successful these methods can be.[4][5] Attractive as these methods are, they are hampered by difficulties when applied to inorganic compounds, especially those containing transition metals. Not only is it the broad span of coordination numbers and coordination geometries accessible to one and the same type of metal which is at the basis of these problems, but it is as well the lack of reliable spectroscopic and thermodynamic data which would allow for an appropriate assessment of force field parameters. Nevertheless, inorganic chemists have successfully addressed conformational problems by force field methods[6][7][8] and have devised different strategies to overcome the intrinsic difficulties. These methods rely by and large on an educated guess of the relevant parameters and the subsequent improvement of these initially chosen parameters by trial and error. The quality of the parameters is generally judged by the accuracy with which the chosen set of parameters will reproduce the structures of a few selected examples of the class of compounds to be modelled. There are two basic problems associated with this type of methodology:

(a) The structures which are used to assess the validity of a specific force field approach are taken from solid-state structure determinations. The approach of fitting force field

  1434−1948/98/0606−0693 $ 17.50+.50/0

parameters to these structures is acceptable as a tool of parameter evaluation only if it is shown that these structures and the conformations represented by them reflect the "inner" molecular potential of the compounds and are not significantly influenced by the outer potential acting on the molecular entities within a crystal.

(b) The method of just selecting a few structures and then fitting the force field parameters on this reduced set of information by trial and error procedures does not appear to be the state of the art with respect to the computing capabilities now available.

In an attempt to understand the conformational flexibility of *tripod*−metal templates, an alternative strategy has been developed, which tries to make the maximum use of the information contained in molecular structures as determined by solid-state crystallography.

With regard to the problem outlined under (a) above, a statistical analysis of 82 structures containing the *tripod*−metal template $CH_3C(CH_2PPh_2)_3Co$ in compounds of the type *tripod*−CoL_2 and *tripod*−CoL_3 was performed in order to find out whether these structures might be taken as an unbiased sample representing the inner molecular forces, not largely disturbed by the crystal environment.[9][10] Irrespective of the statistical tools used (factor analysis, cluster analysis, partial least-squares, scattergraphs, neural networks), the analysis showed that the observed conformations are determined by the inner forces, and thus that they correspond to local minima on the molecular energy hypersurface. This result means that the conformations of *tripod*−metal compounds as determined by X-ray crystallography contain the information necessary for developing an appropriate force field: it is a necessary condition for a set of force field parameters to reproduce these conformations as at least local minima on the energy hypersurface described by it.

With regard to point (b) above, the task of deriving parameter sets from the observed conformations that give the optimal reproduction of the observed structures is a problem of global optimization. This class of mathematical problem has not yet found a unique solution in applied mathematics, nor is there a solution guaranteeing that the global minimum will be found in a finite number of steps in general. Nevertheless, several types of approaches exist, the practicability of which has been demonstrated in quite a number of optimization problems.[11][12] In the present work, Genetic Algorithms (GA) have been used as the optimizing tool. In this paper, we report how, by use of this approach, sets of force field parameters may be automatically derived from a structural database. Compounds of the type *tripod*−$Mo(CO)_3$ [*tripod* = $RCH_2C(CH_2X)(CH_2Y)(CH_2Z)$, X,Y,Z = PR′R″] serve as the specific example. The quality of the parameter sets thus derived is assessed by their ability to reproduce conformations as well as general conformational patterns. It is further validated by the correct prediction of the conformation of a specific *tripod*−$Mo(CO)_3$ compound that had not been part of the data basis.

Data Basis

Compounds of the type *tripod*−$Mo(CO)_3$ containing *tripod* ligands $RCH_2C(CH_2X)(CH_2Y)(CH_2Z)$ with up to three different donor groups X,Y,Z = PR′R″ are generally accessible by reacting $(CH_3CN)_3Mo(CO)_3$ with the appropriate *tripod* ligand.[13][14][15][16] The structures of a number of such compounds have been determined by X-ray crystallography.[13][14][15][16] When this work was commenced, nine relevant structure determinations had been reported[17] referring to ten crystallographically independent molecules of this type. To illustrate the overall structural characteristics of this type of compound, Figure 1 shows two mutually orthogonal projections of the structure of $CH_3C(CH_2PPh_2)_3Mo(CO)_3$.

Figure 1. Two orthogonal projections of the solid-state structure of $CH_3C(CH_2PPh_2)_3Mo(CO)_3$; the numbering scheme given is used throughout the text

The constitutions of all the structural examples included in the data base are given in Table 1.

In order to elaborate a force field description for this class of molecules, the parameters involving bonds to the molybdenum atom had to be evaluated from first principles.The force constants describing the organic part of the compounds on the other hand were taken from the well-established MM2* force field.[4][18] The force field parameters describing the force field interactions involving the metal were incorporated as defined by the expressions given in Figure 2. These parameters were in part refined by the methods described below.

Refinement of the Force Field: Any refinement of the force field will correspond to a search for the set (or sets) of parameter values that best reproduce the conformations present in the data base as local or even global minima on the corresponding energy hypersurface (surfaces). To perform this search in an automatic way, the strategy of Genetic Algorithms has been used throughout in this work. In the realm of molecular modelling, Genetic Algorithms have already been used to scan an energy hypersurface for minimum energy conformations. In the analysis and prediction of the conformation of biological macromolecules, the introduction of these algorithmic methods has in some cases dramatically improved the validity of the predictions.[19][20] To the problem of conformational search on an energy hypersurface, an inverse problem exists: given the confor-

Table 1. Compounds **1**−**10** and rms deviations between the conformations observed and calculated for the two different parameter sets *mm2f* and *mm2t*, and rms deviation between the conformations as calculated on the basis of the two different parameter sets. Compounds **6** and **7**, labelled with an asterisk, are two crystallographically independent conformations of $CH_3C(CH_2PEt)_3Mo(CO)_3$ as found in one and the same unit cell[a]

number	formula	RMS_{mm2f} [Å]	RMS_{mm2t} [Å]	$RMS_{mm2f/mm2t}$ [Å]
1	$CH_3C(CH_2PPh_2)_3Mo(CO)_3$	0.286	0.248	0.060
2	$CH_3C(CH_2PBzlPh)_3Mo(CO)_3$	0.417	0.357	0.190
3	$CH_3C(CH_2PMe_2)_3Mo(CO)_3$	0.197	0.190	0.071
4	$PhCOOCH_2C(CH_2PMe_2)_3Mo(CO)_3$	0.585	0.581	0.061
5	$CH_3C(CH_2PNap_2)_3Mo(CO)_3$	0.353	0.328	0.090
6*	$CH_3C(CH_2PEtPh)_3Mo(CO)_3$	0.252	0.240	0.075
7*	$CH_3C(CH_2PEtPh)_3Mo(CO)_3$	0.243	0.193	0.144
8	$ClCH_2C(CH_2PPh_2)_3Mo(CO)_3$	0.235	0.224	0.059
9	$CH_3C[CH_2P(DBP)_2]_3Mo(CO)_3$	0.312	0.284	0.062
10	$CH_3C(CH_2PEt_2)_3Mo(CO)_3$	0.248	0.235	0.074
mean		0.313	0.288	0.089

[a] Abbreviations: Et: ethyl, Me: methyl, Bzl: benzyl, Nap: naphthyl, DBP: dibenzophospholyl.

Figure 2. Definition of parameters involving the metal center

[a] Colour codes of atom types shown for a fragment of **2** (see Table 1). − [b] Parameter types and corresponding potential terms applied in the force field. − Abbreviations: E: energy; k_b: bonding constant; r: bond length; r_0: equilibrium bond length; k_a: force constant for angle bending; α/α_0: angle/equilibrium angle; $k_{t(1-3)}$: force constant 1−3 for torsion; ω: dihedral angle.

mations which correspond to − at least local − minima on the energy hypersurface, a search on the hypersurface defined by the force field parameters themselves should lead to an optimized parameter set.[21] With the efficiency of Genetic Algorithms in mind, and with the given formal logical analogy of these two problems, it appears natural to apply Genetic Algorithms to adapt a force field to a set of conformations. The formal logical complementarity of the two problems, (a) to find an optimized conformation on the basis of a given force field, and (b) to find an optimized force field on the basis of given conformations, is illustrated in Figure 3.

Figure 3. Complementary problems − analogous solutions

[a] Adaptation of a conformation to a given force field. − [b] Adaptation of force field parameters to a given conformation

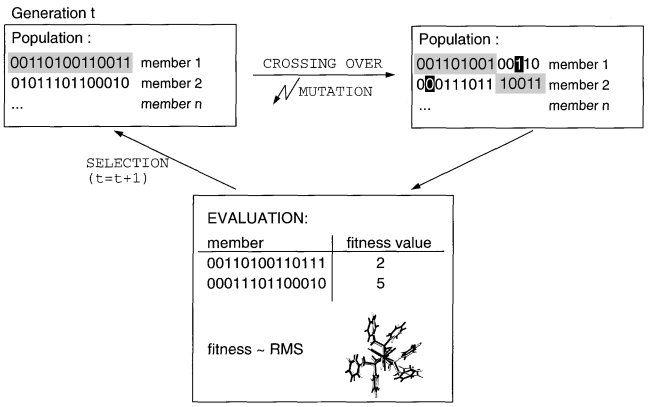
In case (a), the conformation − symbolized by a circle, a square and a triangle, connected in a gear-like arrangement − has to be changed such that it optimally fits into a given pre-designed matrix, the structure of which is strictly defined by the force field used. In case (b), on the other hand, the conformation of the object is fixed and it is the matrix which has to adapt its shape to the conformation presented by the object, i.e. the force field parameters defining the shape of the matrix have to be changed to this end. To solve

these problems in both cases, even if they are logically complementary an optimization procedure must be invoked.

To devise such a procedure, whatever form it takes, the relative quality of a given approximation has to be defined by some numerical qualifiyer that numerically represents the overall performance of this particular approximation. In terms of the language of Genetic Algorithms, this qualifier is called the "fitness value" of a given approximation. In the case considered, the root-mean-square deviation between observed and calculated atom positions is a simple and efficient qualifier and is therefore used throughout in this work. An optimization procedure, whatever its form, has then to find approximations which minimize the discrepancy between observation and model, i.e., in the language of Genetic Algorithms, to find approximations with the optimal fitness values. Because of the a priori unpredictable shape of the hypersurface on which the minimization procedure is performed, analytical methods based on different types of gradients will not generally lead to the desired solution. Purely stochastic methods, on the other hand, will in principle solve the problem, albeit possibly only in an infinite amount of time. With these Monte Carlo type procedures, the hypersurface is probed at random and finding a solution basically relies on Cicero's old statement: "Quis est enim, qui totum diem iaculans, non aliquando conlineet ?"[22] Imagine how long it might take to find the highest elevation in the Alps by this purely stochastic procedure.[23] While completely deterministic analytical methods cannot generally solve the problem and while, as described, stochastic methods may be practically inapplicable, a combination of deterministic and stochastic strategies may well be successful.

Genetic Algorithms may be seen as an algorithmic implementation of both these principles.[24][25][26] The terminology associated with Genetic Algorithms is by and large the same as that used in biology to describe the evolutionary process. As far as we know, the Darwinian principle of "survival of the fittest" is based on two underlying basic molecular processes: crossing over and mutation. The higher the fitness of an individual within a given population, the higher its chance of reproducing; the offspring resulting from sexual reproduction will then have DNA blueprints from the mother as well as from the father, which will generally have fitness values well above the average. In this way and by occasional mutations, sufficient diversity is retained in each generation of a population such as to guarantee dynamic evolution. On the other hand transmithose DNA blueprints which have already led to a high fitness of the parent individuals is at the basis of the evolutionary optimization process and represents the deterministic part of it. Pieces of information which have already been evaluated as being efficient are at least in part inherited and used in the next evolutionary optimization step. To apply this evolutionary strategy to numerical problems, a set of parameters for which the optimized values are to be found is encoded as a binary string (Figure 4).

Figure 4. Force field parameter refinement with Genetic Algorithms

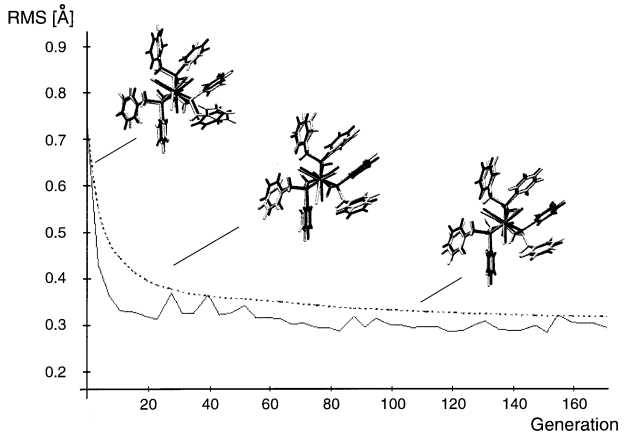


A "population" of such strings, each one representing a set of parameters, is generated at random. The fitness of all these parameter sets is then evaluated. For the problem at hand, this means that the set of parameters is applied to optimize the structures starting with the conformations as represented in the database. The rms deviations (see Experimental Section) between the structures as derived from the force field approach and those in the database are calculated, and taken to assess the fitness value for this specific set of parameters. This procedure is performed for all the members of a population (Figure 4). To produce the next generation with an intentionally increased fitness, those members of the parent population that are already ranked highly on the fitness scale are preferentially selected to reproduce. Mating is simulated by switching the information encoded in the binary strings, with the switching point being selected at random (Figure 4). By this crossover procedure, the population of the next generation is built up. To increase diversity, point mutations − changing a 0 to a 1 and vice versa − are occasionally performed at random. The procedure is then repeated to produce generation after generation and it is generally observed that the fitness increases from population to population. A typical result of the application of this procedure to the problem at hand is shown in Figure 5.

The dotted line represents the mean of the fitness values of all precedent populations at a given generation. It is seen that the discrepancy decreases exponentially. The lower curve represents the overall fitness of each individual population. It is seen that even after 160 generations, the fitness is still able to change. The population is not stuck in a niche, but is still capable of adapting. There is no guarantee that the optimal set of parameters has already been found, nor is there of course, by the very nature of the problem, any guarantee that a unique solution exists at all. The increase of the quality with which the best parameter set in a generation reproduces the conformation in the database is shown in Figure 5 as an overlay of the computed (grey) and the observed (red) conformations of **2** as an example.

A great advantage of applying Genetic Algorithms as minimizers stems from the fact that their algorithmic struc-

Figure 5. Genetic Algorithms at work[a]



[a] The force field model is optimized such that the calculated conformations (i.e. grey) will reproduce the observed ones (i.e. red) increasingly well. The dotted line represents the cumulated fitness, which steadily increases as indicated. The average fitness of each individual population is represented by the continous line.

ture is inherently parallel, and hence the time-consuming steps (in the present case force field minimization of 10 structures) may naturally be distributed over the equivalent number of processors in a parallel computing environment (see Experimental Section). Yet another advantage may be seen in the fact that Genetic Algorithms have no problems in evaluating a whole subspace of solutions that will reproduce the observations equally well. Why should there exist a subspace of solutions and not just one optimal solution to a given problem? In the problem at hand, it is in fact to be expected that there will be several combinations of force field parameters, which − although individually numerically quite different − will reproduce the observations in conformational space equally well. This is to be expected owing to the fact that there are strong and sometimes even linear dependencies between the parameters. Consider a triangle with sides *a*, *b*, *c*. Given the length of the sides *a* and *b*, the length of *c* is as well determined by the angle subtended between *a* and *b* as it is by the value for the distance itself. In a force field approximation as applied the angle as well as the distance would be determined by individual potential functions for the angle and its deformation on the one hand and the distance and its change on the other. Trying to extract force constants for these two functions from a set of observations of triangles would inevitably lead to a multitude of solutions − in principal an infinite number of solutions − corresponding to the linear dependence between the angle and the distance. All these solutions will, of course, reproduce the observations equally well. While in spectroscopy this problem has been well delineated and is appropriately handled[27], molecular modelling as generally applied does not bother about it. As far as the refinement of force field parameters on the basis of a given set of conformations is concerned, the dependencies in parameter space would be cumbersome to handle by analytical refinement methods, while Genetic Algorithms would just produce a collection of solutions that would all reproduce the observations equally well.

With regard to the physical meaning of the parameters derived by such a refinement, it is evident from the above that individual parameters as such do not necessarily have to have a definite physical meaning. It is only the whole ensemble of parameters, which altogether have the physical meaning that, fed into the appropriate force field program, they will reproduce and hopefully predict conformations in agreement with empirical observations. Nevertheless, it is pleasing to apply, wherever possible, force constants that have been assessed by independent physical methods, and which then have a definite meaning even as individuals. With this in mind, the $Mo-C_{CO}$ stretching constant was set throughout at a value taken from the literature (Table 2), as was the Mo−C−O bending force constant.[28][29] Two parameter sets were finally refined (Table 2), with different degrees of freedom. The set designated as *mm2f* has the two aforementioned force constants at fixed values. The equilibrium distances and angles were set at the mean values observed for the sample in all cases, except for the $Mo-C_{CO}$ distance and the Mo−C−O angle, which were allowed to refine while the corresponding force constants were held at the fixed values (Table 2). Refinement thus mainly applied to the different types of force constants (Table 2). In the set of parameters designated as *mm2t*, the above set was augmented by allowing the torsion potentials involving contributions from molybdenum to refine (Table 2). This set was created for several reasons. Firstly, it seemed of interest to ascertain whether an augmented set of parameters would lead to a better refinement, as would in principle be expected, and whether the refinement procedure as applied would be able to converge on a set of values, even though these parameters are not so well-defined by the variance apparent in the data set. (The variance of values referring to these parameters in the data set would not lead one to expect that a stringent refinement would be possible at all). The second reason for deriving an augmented set of parameters was the curiosity to see how the introduction of these additional potentials would influence the values of the force field parameters common to both parameter sets. It is seen that the individual values for bond stretching and angle bending parameters are indeed quite different in both sets.

If these two sets of parameters are now applied to the ten compounds in the data base, to find the local minimum that best corresponds to the observed conformation, the results obtained (Table 1) are very similar for both parameter sets, even though individual parameters are grossly different in the two cases (Table 2). The rms deviation between observed and calculated structures as defined above is consistently somewhat smaller with the augmented force field *mm2t* than that derived using the *mm2f* approach (Table 1). This applies to the individual structures as well as to the sample as a whole (Table 1). With only a small difference between the quality of fit obtained from force field calculations based on the *mm2f* parameter set and that based on the *mm2t* set, the latter having 15 additional parameters, it

appears that on statistical reasoning the set *mm2f* is built on firmer grounds and should thus be preferred. Anyhow, the rms deviation between the two conformations for each compound obtained by applying one or the other set in the force field geometry optimization is considerably smaller than the rms of the difference between the observed and calculated structures in each case (Table 1). To illustrate the quality of fit obtained from the parameter set *mm2f*, Figure 6 shows an overlay of experimental and calculated conformations of four examples taken from the data set.

Figure 6. Overlay of calculated (grey) and observed (red) structures of compounds **1**, **3**, **5** and **6** (see Table 1)

The agreement is clearly quite satisfying. It may well be that the predominance of (mainly repulsive) interactions between the organic moieties of the ligands, which is characteristic of *tripod*-metal templates, is largely responsible for this degree of agreement. The conformations are thus dominated by forces that have been thoroughly modelled for organic compounds.[4][5]

If the parameters as evaluated here have the physical meaning of reproducing the conformational behaviour of *tripod*-metal compounds as a whole, the general trends and regularities observed for these compounds must be reliably reproduced by the force field approach. By analysis of 82 *tripod*-metal compounds[9], it has been found that the torsional positions occupied by the phenyl groups of the PPh_2 donors of *tripod* ligands form a regular pattern when plotted against each other in a type of scattergraph.[9] This analysis is shown in Figure 7 (black dots).[30]

Figure 7. Contour plot representing lines of equal energy as calculated for compound **1** based on the *mm2f* parameter set, with relative energies colour-coded as shown. The coordinates (φ) refer to the rotational positions of the two phenyl groups of a PPh_2 group in **1**. The black dots represent experimental points based on the analysis of 82 structures.[9] Pictograms of torsion angles φ refer to a projection of the molecule as presented in Figure 1 (left). For each PPh_2 group, φ_{2n+1} and φ_{2n} follow each other when counting in a clockwise sense referring to this projection. Grey lines represent energetically feasible pathways for rotational reorientation

This scattergraph is projected onto a contour-line diagram representing lines of equal energy as computed on the basis of the *mm2f* parameter set. It is evident from Figure 7 that regions of low energy (1, 2) correspond to regions which are densely occupied by experimental points. Regions of high energy (6) are not occupied at all. Region 3 corresponding to a two-ring flip transition[9] is in contrast quite densely populated. The contour plot shows that the energy differences between the regions 1 and 2 on the one hand and 3 are only small. Regions 4 and 5 correspond to one-ring flip transitions.[9] It has already been argued[9] that the extension of experimental points towards these regions indicates the feasibility of one-ring flip transitions. The potential pathways for a rotational rearrangement of the phenyl

rings at a PPh_2 grouphad already been delineated by lines as shown in Figure 7[9]. It is satisfying to see that the energy separation between the regions 3 and 2 and the transition point 4 is modest (roughly 10 kJ/mol), so that the free rotation of such phenyl groups as observed by NMR even at a temperature of −80°C finds is counterpart in the diagram.

Satisfying as the result shown by Figure 7 is, a more specific validation of the force field approach as developed would be the prediction of the conformation of a specific compound that had not already been included in the data set used to derive the force field. For this purpose, the conformational space available to $CH_3C[CH_2P(o\text{-Tol})_2]_3$-$Mo(CO)_3$ (**11**) was systematically analyzed by the above force field approach. It was not clear at that time that the compound could be prepared and even structurally characterized by X-ray analysis, and so it was especially satisfying to find that the subsequently determined crystal structure[31] agreed exceedingly well with the structure obtained as the global minimum from the force field calculations.

Figure 8. Overlay of the conformation of $CH_3C[CH_2P(o\text{-Tol})_2]_3$-$Mo(CO)_3$ as observed by X-ray analysis and calculated on the basis of the *mm2f* set (global minimum)

Figure 8 shows an overlay of the conformation as obtained from the conformational search (grey) and that observed in the crystal (red). Irrespective of the parameter set used (*mm2f*, *mm2t*), the global minimum was found at the same place in conformational space. The rms deviation between the two model conformations amounted to only 0.11 Å. The rms difference between the model conformations and the conformation observed in the crystal is very small as well (*mm2f*: 0.37 Å, *mm2t*: 0.42 Å)

The results obtained to date using this novel type of approach are sufficiently promising to warrant further investigations along such lines. While, from the results presented, it appears probable that the conformations of related molecules will be reliably predicted, it will be interesting to see whether the calculated energy hypersurfaces are approximately on scale with experimental data. The application of this method to problems pertaining to the conformational flexibility of *tripod*−$Mo(CO)_3$ compounds in solution (forthcoming paper) should give an answer to this question.

In summary, a novel approach to the derivation of force field parameters for metal-containing compounds is presented. It is shown that refinement of force field parameters on the basis of solid-state structures can be efficiently performed by the application of Genetic Algorithms. These algorithms are robust and allow for the simultaneous refinement of many force field parameters on an extended data basis containing many structures. When applied to *tripod*−$Mo(CO)_3$ compounds, the approach leads to models with a high predictive power with respect to specific conformations of these molecules, as well as to their overall conformational behaviour.

Financial support by the *German Science Foundation* (DFG HU 151/24-1), the *Fonds der Chemischen Industrie* and the Graduiertenkolleg "*Selektivität in der organischen und metallorganischen Synthese und Katalyse*" is gratefully acknowledged. The computations were made possible by grants of computing time at *Parsytec (IWR Heidelberg*) and CRAY (*KFA Jülich*) parallel computing environments. One of us (J. H.) wishes to thank the Graduiertenkolleg "*Modellierung und wissenschaftliches Rechnen in Mathematik und Wissenschaft*" (*IWR Heidelberg*).

Experimental Section

(a) *Force Field Calculations:* In order to apply the methodology of Genetic Algorithms to the problem of refining sets of force constants, it is necessary to have the force field program available as a source text. Since almost all modern force field programs are nowadays only commercially available and are delivered to the customer as executable files that will run only on a specific type of machine, these commercial programs are of no help. Instead of writing a whole new package for force field calculations, YAMMP[32], a program set written by R. K. Z. Tan et al. in ″C″, which is available as a public domain source code, was used as the core. The force field expressions implemented in this program were changed to correspond to the potentials applied in MacroModels MM2*.[18] Test runs on a vast sample of different compounds were performed with both programs to ensure that the results produced by MacroModel and those produced by the modified YAMMP program were numerically identical. Two converters which transform the MacroModel data format into the YAMMP data format were implemented and YAMMP was embedded in a number of shells so as to allow its use on different types of single processor machines (e.g. Pentium, SGI) as well as on different types of parallel computing systems (Parsytec GC, CRAY T3E).

To evaluate the rms deviation between two structural models, the mutual translational and rotational positions of the two models were optimized by a least-squares procedure using in part the *quatfit* program of D. Heisterberg[33] and in part a program based on the same algorithmic approach written by K. Allinger. The root-mean-square deviation was defined as the square root of the sum of the squares of the deviations between all n corresponding pairs of atoms divided by the square root of n (hydrogen atoms were included in these calculations throughout).

(b) *Optimization by Genetic Algorithms:* The force field parameter refinement was performed with the program *GAPAO*, written in ″C″, which uses functions from the *PGA-Genetic Algorithm Package.*[34] These functions are fully parallelized and are based on the *Message Passing Interface Library.*[35] This standard parallelization platform is available for all types of Unix systems, even for a net-

work of Linux PC's, and thus *GAPAO* has been installed on a Parsytec GC/Power Plus as well as on a CRAY T3E. Parallelization in this case means that all the evaluations of the parameter sets − implying the time-consuming energy minimizations of all the basis conformations − are performed on different processors. This leads to a tremendous economization of time and is thus another important advantage of the optimization procedure as applied.

Different crossing-over types and different selection routines are available.[34] In the cases reported, the approach of uniform crossing-over has been used.[36] To this end a mask of 0s and 1s with the length of the corresponding bitstring is generated at random. Applied according to the crossing over probability this mask then indicates which bits from the parents have to be taken to produce offspring. This type of crossing-over turned out to be especially useful in the treatment of large-scale parameterization problems.

The traditional "fortune wheel" selection type has been used as the selection routine.[25] This type of selection means that the probability of a given bitstring to be chosen to produce an offspring is proportional to its fitness value. Furthermore, an elitistic optimization strategy[25] has been adopted throughout, i.e. the best parameter set found in a given population is always left unchanged and taken into the next generation.

Evaluation was performed by energy minimization of all the structures contained in the data basis. To this end, the function *KA MMOPTfct*, written in C by K. Allinger, was used. The Polak-Ribière conjugate gradient minimization method[11][12] was employed and minimization was always driven to convergence (convergence criterion: 0.01 Å/iteration).

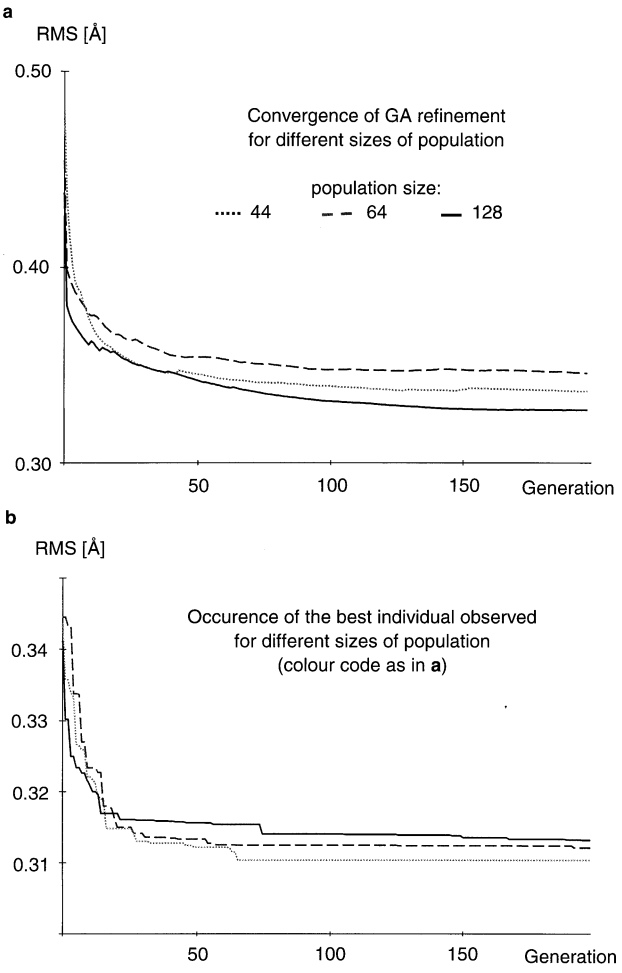
Parameters referring to force constants for bond stretching were encoded in the range 0.0−2.0 mdyn/Å. Parameters referring to force constants for angle bending were encoded in the range 0.0−1.5 mdyn/(rad^2). Torsional force constants k_{t1}, k_{t2}, and k_{t3} (see Figure 2) were encoded in a range between −1.5 and +1.5 kcal/mol. The refined Mo−C−O angle (see Table 2) was optimized in an interval between 170° and 180°; the Mo−C bond length was allowed to refine in an interval from 1.7 Å to 2.2 Å. Different subsequent runs were performed, each taking as its starting point the best results of the foregoing process, gradually augmenting the number of bits encoding the corresponding parameters in a stepwise manner from 5 to 8, and thus improving the resolution of the refined parameters.

Another way of overcoming the problem of discreteness in the optimization process was to apply a "hill climbing heuristic" to the best parameter set found and thus to locally optimize the corresponding parameters. This feature is implemented in the *GAPAO* program (*GAPAO-h)* by the strategy of altering or diminishing in a stepwise manner one parameter after another in the best parameter set, such that the resulting rms deviations become smaller. This is done until no further amelioration is achieved by such adjustments.

The method as described lends itself to a principally unlimited number of variations and it is a matter of experience to select appropriate values for the size of a population, and for the rates of crossing-over and mutation. This is exemplified in Figures 9a and b, which refer to different refinement runs for the parameters characterized by *mm2f*.

The rates of crossing-over and mutation were left unchanged for the three runs shown in Figure 9, with values that were found to be especially appropriate for the parameterization problem at hand (values given in Figure 9). The size of the population was different for the three runs. Not unexpectedly, the cumulated overall fitness,

Figure 9. How to find the appropriate GA protocol. Parameters used in the GA runs presented in a and b: crossing-over rate = 0.8, mutation rate = 0.02



which is the mean of all the previously assessed fitness values, improves with the size of the population (see Figure 9a). However, the real aim of the optimization is not to find the population with the maximum overall fitness, but rather to find one member of a population − from which ever generation − which is the fittest individual of all, i.e. the search is aimed at the best set of force field parameters. Figure 9b illustrates that of the three runs described, the one based on a population of 60 randomly selected parameter sets produces the optimal solution after about 60 generations, with no better solution being found after more than 100 subsequent generations. The run based on a population size of 128 obviously takes far longer to converge. The run based on 44 individuals shows rapid convergence, but appears to be stuck at a lower quality solution.

In summary, it may be said that quite some experience and in some respects "green fingers" are needed to successfully apply Genetic Algorithms. However, this statement applies to other methods of global optimization as well and is not specific to the physical model analyzed here.

Calculations were performed on a CRAY T3E-512 with 512 compute nodes (600 MFLOPS/processor, 128 MB/processor, 333 MHz) and a Parsytec GC/Power Plus-192 (192 Power PC 601, 80 MHz, 32 MB/processor).

(c) *Conformational Analyses:* The φ_1/φ_2 contour plot of the PPh_2 conformations in **1**, as shown in Figure 7, is based on a grid search with a resolution of 5° for each P−Ph rotation φ (a φ value of 0° corresponds to an orientation of the phenyl ring perpendicular to the plane defined by the three P atoms, see Figure 7). Due to the C_2 symmetry of the phenyl groups, their rotation is periodic with a period of 180°, thus leading to 36 × 36 = 1296 grid points to describe one PPh_2 group. To account for the molecule as a whole, the multitude of orientations accessible to the remaining PPh_2 groups and to the chelate scaffolding have to be appropriately represented in the corresponding calculations. To this end, the torsion of the scaffolding τ[9] was uniformly set to starting values of 11° or 34°, respectively, leading to 2 × 1296 = 2592 starting geometries. In these conformations, the rotational arrangement at the remaining PPh_2 groups was set such that it corresponded to the regularites derived from an extensive analysis of such compounds ($\tau_{1/2/3}$ = 11°: $\varphi_{3/5}$ = 36°; $\varphi_{4/6}$ = 49°; $\tau_{1/2/3}$ = 34°: $\varphi_{3/5}$ = −50°, $\varphi_{4/6}$ = −29°; for the meaning of the numbers, see Figures 1 and 7).[9] (The conformations of a molecule such as **1** are chiral in general; only one of the enantiomers needs to be considered in the grid search due to this mirror symmetry in conformational space.)

A total of 2592 starting geometries were generated by the program *confgenrub* (written in "C" for that purpose) and then minimized by the program *bmin* as included in Version 5.0 of the molecular modelling software package MacroModel[18], using the force field MM2* and the parameter set *mm2f* described above. In order to evaluate the energy at each of the 1296 grid points, the rotational positions at the selected PPh_2 group at P1 (the orientation of the phenyl group as defined by the coordinate values of the grid point) had to be fixed during minimization. This was achieved by the technique of using a dummy atom, Du: A dummy atom was placed at a distance of 1 Å from the atom P_1, perpendicular to the plane defined by the 3 phosphorous positions, in the direction pointing towards the metal atom. Technically, this was achieved by imposing 3 restraints on its position: the distance P_1−Du as well as the angles P_2−P_1−Du and P_3−P_1−Du were fixed by giving them suitably steep potentials. The torsion angles φ_1 and φ2 were kept at the preordained values by imposing a corresponding restraint on the relevant torsion angles Du−P_1−C_{ipso}−C_{ortho}. Force constants as applied for this purpose were: distances: 500 kJ/Å^2, angles: 9999 kJ/rad^2, torsion angles: 9999 kJ/mol for k_{t1} (see Figure 2).

The refinement was performed using the Polak-Ribière minimizer with a gradient of 0.01Å/iteration and a maximum of 500 iteration steps as the limiting criteria. No cut-off was set for nonbonded interactions.

After minimization of each pair of conformations (two different τ values, see above) calculated for every grid point, the one with the lower energy was chosen to represent the energy (diagram in Figure 7).

The final contour plot as shown in Figure 7 was obtained by taking into account the isoenergetic enantiomeric conformation in each case. The mirror symmetry of the problem is apparent by a diagonal mirror plane characterizing the symmetry of the plot. The *plot2D* option of MacroModel was used to prepare the contour plot shown in Figure 7. The result of this procedure as depicted was obtained using the *mm2f* parameter set. Repeating the same procedure with the *mm2t* set produced a contour plot with the same overall characteristics.

A complete conformational search was performed for **11**. To generate the starting conformations for the search, the rotations of the tolyl groups about their P−C_{ipso} bonds were set at the initial values of φ = −135°, −45°, 135°. These torsional positions were defined with respect to a dummy atom, the position of which was generated as described above. The φ values refer to the torsion angle Du−P−C_{ipso}−$C_{ortho(Me\text{-}substituted)}$. The torsional arrangement of the cage was set by assigning a value of 20° to each of the 3 torsion angles τ.

Combining the four predefined orientations of each tolyl group in all possible combinations for the six tolyl groups present in the molecule gives a total of 4^6 = 4096 starting conformations. In this set, there are 16 (2^4) conformations that occur only once because of the C_3-symmetry of the molecule. The remaining 4096 − 16 = 4080 conformations occur in isoenergetic triplets. While physically indeterminable, they differ in assigning the same values of the six tolyl rotations φ to the P(*o*-Tol)$_2$ groups at P1P2P3, P2P3P1 or P3P1P2 in this sequence of cyclic permutations. Thus only (4096 − 16)/3 + 16 = 1376 starting geometries had finally to be taken into account. Minimization was performed as described above, with the criteria of a limiting gradient of 0.001Å/iteration and the maximum number of 2000 iterations allowed.

The calculations were carried out on two Silicon Graphics Indigo^2 MIPS R4400 workstations, 200 MHz, 128 MB RAM, operating under IRIX 5.3.

Table 2. Parameter sets *mm2t* and *mm2f*. Refined parameters are labelled with an asterisk

(a) Bond stretching

	parameter set *mm2f*		parameter set *mm2t*	
	r_0 [Å]	k_b [mdyn/ Å]	r_0 [Å]	k_b [mdyn/ Å]
Mo−C1	1.99*	2.00	1.94*	2.00
Mo−P	2.53	1.98*	2.53	0.75*

(b) Angle bending

	parameter set *mm2f*		parameter set *mm2t*	
	α_0 [°]	k_a [mdyn/rad^2]	α_0 [°]	k_a [mdyn/rad^2]
P−Mo−P	83.70	1.18*	83.70	0.53*
C1−Mo−C1	87.10	0.33*	87.10	0.39*
C1−Mo−P	94.90/175.90	0.1*	94.90/175.90	0.25*
Mo−C1−O	173.0*	0.47	170.00*	0.47
Mo−P−C2	119.0	0.29*	119.00	0.43*
Mo−P−C3	114.50	0.02*	114.50	0.29*

(c) Torsion

	parameter set *mm2f*			parameter set *mm2t*		
	k_{t1}[a]	k_{t2}[a]	k_{t3}[a]	k_{t1}[a]	k_{t2}[a]	k_{t3}[a]
P−Mo−P−C3	0.00	0.00	0.00	−0.90*	−0.58*	0.71*
P−Mo−P−C2	0.00	0.00	0.00	1.05*	0.26*	−0.24
Mo−P−C3−C3	0.00	0.00	0.00	−0.37*	−0.55*	0.11*
Mo−P−C2−C2	0.00	0.00	0.00	−1.06*	−0.19*	1.14*
Mo−P−C3−H	0.00	0.00	0.00	0.72*	−0.92*	1.20*

[a] In kcal/mol.

[1] See for example: [1a] H. Brintzinger, D. Fischer, R. Muhlhaupt, B. Rieger, R. M. Waymouth, *Angew. Chem. Int. Ed. Engl.* **1995**, *34(11)*, 1143−1170. − [1b] C. R. Landis, C. R. Halpern, *J. Am. Chem. Soc.* **1987**, *109*, 1746−1754. − [1c] G. Helmchen, S. Kudis, P. Sennhenn, H. Steinhagen, *Pure Appl. Chem.* **1997**, *69*, 513−518.

[2] G. Frenking, I. Antes, M. Boehme, S. Dapprich, A. W. Ehlers, V. Jonas, A. Neuhaus, M. Otto, R. Stegmann, A. Veldkamp, S. F. Vyboishchikov, in *Reviews in Computational Chemistry*, *Vol. 8* (Eds.: K. B. Lipkowitz, D. B. Boyd), VCH Publishers, Inc., New York, **1996**, p. 63−130

[3] L. J. Bartolotti, K. Flurchick, in *Reviews in Computational Chemistry*, *Vol. 7* (Eds.: K. B. Lipkowitz, D. B. Boyd), VCH Publishers, Inc., New York, **1996**, p. 187−259.
[4] N. L. Allinger, *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.
[5] U. Burkert, N. L. Allinger, *Molecular Mechanics*, ACS monograph, Washington, **1992**.
[6] P. Comba, T. W. Hambley, *Molecular Modelling of Inorganic Compunds*, VCH, Weinheim, **1995**.
[7] B. J. Hay, *Coord. Chem. Rev.* **1993**, *126*, 177−236.
[8] C. R. Landis, D. M. Root, T. Cleveland, in *Reviews in Computational Chemistry*, *Vol. 6* (Eds.: K. B. Lipkowitz, D. B. Boyd), VCH Publishers, Inc., New York, **1995**, 73−148.
[9] S. Beyreuther, J. Hunger, G. Huttner, S. Mann, L. Zsolnai, *Chem. Ber.* **1996**, *129*, 745−757.
[10] G. Huttner, S. Beyreuther, J. Hunger, in *Software-Entwicklung in der Chemie 10* (Ed.: J. Gasteiger), GDCH, Frankfurt am Main, **1996**, 201−207.
[11] W. H. Press, *Numerical Recipes: The Art of Scientific Computation*, Cambridge University Press, Cambridge, **1986**.
[12] P. E. Gill, W. Murray, M. H. Wright, *Practical Optimization*, Wiley, New York, **1988**.
[13] O. Walter, Th. Klein, G. Huttner, L. Zsolnai, *J. Organomet. Chem.* **1993**, *458*, 636−640.
[14] A. Muth, O. Walter, G. Huttner, A. Asam, L. Zsolnai, *J. Organomet. Chem.* **1994**, *468*, 149−163.
[15] Th. Seitz, A. Muth, G. Huttner, Th. Klein, O. Walter, M. Fritz, L. Zsolnai, *J. Organomet. Chem.* **1994**, *469*, 155−162.
[16] Th. Seitz, A. Asam, G. Huttner, O. Walter, L. Zsolnai, *Z. Naturforsch.* **1995**, *50b*, 1287−1306.
[17] The structures **3**, **8** and **10** (see Table1) have not yet been published, but have now been submitted. Details of the crystal structure investigations may be obtained from the Fachinformationszentrum Karlsruhe, D-76344 Eggenstein-Leopoldshafen (Germany), on quoting the depository numbers CSD-406278 (**3**), CSD-406277 (**8**) and CSD-406275 (**10**).
[18] The mm2 force field as implemented in Macromodel 5.0 (mm2*) has been used throughout (see F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, W. C. Still, *J. Comp. Chem.* **1990**, *11*, 440−467).
[19] J. Devillers (Ed.), *Genetic Algorithms in Molecular Modelling*, Academic Press, San Diego, **1996**.
[20] R. Judson, in *Reviews in Computational Chemistry*, *Vol. 10* (Eds.: K. B. Lipkowitz, D. B. Boyd), VCH Publishers, Inc., New York, **1997**, 1−73.
[21] The idea of applying global minimizers to the inverse problem of finding optimized force field parameters has also been mentioned by van Gunsteren: − [21a] P. Ulrich, W. R. P. Scott, W. F. van Gunsteren, A. E. Torda, *Protein Structure Prediction Force Fields: Parameterization With Quasi-Newtonian Dynamics*; *Proteins* **1997**, *27*, 367−384. In contrast, local minimizers have routinely been used to locally improve parameters − [21b] S. Lifson, A. Warshel, *J. Chem. Phys.* **1968**, *49*, 5116−5128.
[22] Cicero, *De divinatione*, II, 59, "If somebody shoots the whole day, shouldn't he occasionally hit?"
[23] The problem is not drastically reduced when instead of the Alps, the Scottish Highlands or the Rocky Mountains are considered.
[24] J. Holland, *Adaption in Natural and Artificial Systems*, MIT Press, Michigan, **1975**.
[25] D. E. Goldberg, *Genetic Algorithms in Search, Research and Machine Learning*, Addison-Wesley, New York, **1989**.
[26] The strategy of simulated annealing is yet another way of combining deterministically analytic and stochastic strategies. See for example: [26a] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **1983**, *220*, 671−680. This methodology has been successfully applied to problems of conformational search, see for example: [26b] S. R. Wilson, W. Cui, J. W. Moskowitz, K. E. Schmidt, *Tetrahedron Lett.* **1988**, *29*, 4373−4376.
[27] E. B. Wilson, J. C. Decious, P. C. Cross, *Molecular Vibrations - The Theory of Infrared and Raman Vibrational Spectroscopy*, Maple Press Company, New York, **1955**.
[28] L. H. Jones, R. S. MacDowell, M. Goldblatt, *Inorg. Chem.* **1960**, *8*, 11, 2349−2363.
[29] Since bending force constants derived from spectroscopy may not be too accurate (see for instance ref.[5]), refinement was also performed including the Mo−C−O force constant in the subset of variable parameters. The force field parameters derived for this augmented set did not show an improvement over those with the bending force constants at the value given. It was observed, however, that the refined values of this force constant tended to be smaller than the value taken from the literature.
[30] In the scattergraph published previously for the same data (ref.[9]), a distortion by a partially incorrect treatment of symmetry had been introduced; this is now corrected in the diagram presented in Figure 7. We thank H. B. Bürgi for helpful discussions.
[31] P. Schober, Ph.D thesis, Heidelberg 1997.
[32] R. K. Z. Tan, S. C. Harvey, *J. Comp. Chem.* **1993**, *14*, 455−470.
[33] D. Heisterberg, **1990**, unpublished results; the program is available from the software archive of the Computational Chemistry List (CCL) at http://ccl.osc.edu/chemistry.html.
[34] D. Levine, Argonne National Laboratory, **1995**; the package is available at http://www.mcs.anl.gov/pgapack.html.
[35] W. Gropp, E. Lusk, Argonne National Laboratory, **1996** (available from http://www.mcs.anl.gov/mpi).
[36] G. Syswerda, *Proc. 3rd Int. Conference on Genetic Algorithms*, Morgan Kaufman Publishing, **1989**.

[97246]